DOI 10.37882/2223-2966.2024.7-2.42

СОВРЕМЕННЫЕ МЕТОДЫ ИСПОЛЬЗОВАНИЯ ИНВЕРТИРОВАННОГО ИНДЕКСА

RECENT METHODS OF USING THE INVERTED INDEX

A. Shrank

Summary. The inverted index algorithm is one of the most popular algorithms used in search engines and enterprise document management systems. The simplicity of the algorithm makes it universal for application to any data representations, which has led to the creation of many adaptations and improvements. Examples of these improvements can be algorithms such as TF-IDF or BM25. The article considers the intrastate and international experience of using the inverted index, its modifications, and ways of adapting to the specifics of recent search engines. The evolution and development vector of the algorithm, the pros and cons of various modifications and their applications are considered.
Now, most search engines use neural language models. This leads to the forced use of embeddings as a way of presenting data. With the analysis of past decisions, a method for applying the inverted index algorithm to neural network embeddings was presented. This solution will allow the use of artificial intelligence where previously it was impossible due to the use of an inverted index, as well as improve search engines using neural network models.

Keywords: inverted index, neural networks, search, search engines, search algorithms, embedding, language models.

Шранк Алексей Александрович
Аспирант, Национальный исследовательский университет ИТМО, г. Санкт-Петербург
alex29shrank@gmail.com

Аннотация. Алгоритм инвертированного индекса является одним из наиболее популярных алгоритмов, используемых в поисковых системах и системах документооборота предприятий. Простота алгоритма делает его универсальным для применения к любым представлениям данных, что послужило созданию множества адаптаций и улучшений. Примерами таких улучшений могут быть такие алгоритмы как TF-IDF или BM25. В статье рассматривается отечественный и зарубежный опыт применения инвертированного индекса, его модификации и способы адаптирования под специфику современных поисковых систем. Рассмотрены эволюция и вектор развития алгоритма, плюсы и минусы различных модификаций и области их применения.
На данный момент большинство поисковых машин используют нейронные языковые модели. Это приводит к вынужденному использованию эмбеддингов как способу представления данных. С учетом анализа прошлых решений, был представлен способ применения алгоритма инвертированного индекса к эмбеддингам нейронной сети. Данное решение позволит использовать искусственный интеллект там, где раньше это было невозможно из-за использования инвертированного индекса, а также улучшить поисковые машины использующие нейросетевые модели.

Ключевые слова: инвертированный индекс, нейронные сети, поиск, поисковые машины, алгоритмы поиска, эмбеддинг, языковые модели.

Введение

Поиск информации, одна из базовых задач которая появилась вместе с понятием информация. Она является неотъемлемой частью нашей повседневной жизни и профессиональной деятельности. В условиях стремительного роста объемов данных, доступных в Интернете и корпоративных системах, эффективность и точность поиска становятся критически важными для пользователей и организаций. Современные поисковые технологии позволяют не только находить релевантные данные за доли секунды, но и анализировать их с целью выявления закономерностей и принятия обоснованных решений.

Инвертированный индекс является одной из наиболее значимых и фундаментальных структур данных в области информационного поиска и обработки текстов. С момента своего появления в середине 20-го века, инвертированные индексы сыграли ключевую роль в развитии поисковых систем, баз данных и других технологий, связанных с управлением и анализом больших объемов информации [1]. Основной принцип работы инвертированного индекса заключается в создании структуры, которая позволяет быстро находить все документы, содержащие определенные слова или фразы, что делает поиск информации гораздо более эффективным по сравнению с традиционными методами последовательного перебора.

Таким образом, углубленный взгляд на историю, принципы работы, развитие и современные применения инвертированного индекса может открыть новые способы на его применение с современными структурами данных, что значительно ускорит их обработку и поиск. Среди таких структур данных могут выступать эмбеддинги нейронных сетей так как их использование в качестве идентификатора данных значительно выросло в связи с ростом использования нейросетевых языковых моделей в поисковых машинах.

Исходя из вышеперечисленного, объектом исследования является алгоритм инвертированного индекса. Предмет исследования — способность адаптации ал-

горитма к новым структурам данных. Цель исследования — рассмотреть возможность использования алгоритма с эмбеддингами нейросетевых моделей.

Литературный обзор

Первые идеи об инвертированном индексе возникли еще в середине 20-го века. В 1950-х и 1960-х годах, когда компьютеры начали использоваться для хранения и обработки текстовых данных, ученые начали разрабатывать методы для быстрого поиска информации. Идея создания списка всех уникальных слов в документе (или коллекции документов) и указания, в каких документах и на каких позициях они встречаются была выдвинута как наиболее эффективная. В 1960-х годах инвертированный индекс стал более формализованным благодаря развитию компьютерной техники. Системы, такие как SMART (System for the Mechanical Analysis and Retrieval of Text) под руководством Джерарда Салтона, были одними из первых, кто использовал инвертированные индексы для информационного поиска [2].

С развитием информационных технологий инвертированные индексы начали активно использоваться в коммерческих продуктах, таких как базы данных и поисковые системы. Улучшения в алгоритмах и архитектуре компьютеров позволили сделать его более эффективным и масштабируемым.

В 1990-х годах с развитием Интернета инвертированные индексы стали основой для веб-поисковых систем. Поисковые системы, такие как AltaVista, Google и другие, использовали инвертированные индексы для быстрой и эффективной индексации и поиска по огромным объемам веб-контента [3].

Google внес значительные улучшения в алгоритмы ранжирования и обработки запросов, использовав инвертированные индексы как часть своей инновационной поисковой технологии. Это стало точкой развития модификаций алгоритма. Условно можно выделить пять направлений:

1. Сжатие индексов — использование методов сжатия данных для уменьшения объема инвертированных индексов, что помогает снизить затраты на хранение и ускорить доступ к данным [3].
- Гамма-кодирование
- Elias-кодирование
- Ро-Лекка кодирование

Это наиболее популярный метод повышения эффективности так-как возможности его применения ограничиваться только представлением самих индексов.

2. Преобразование слов.
- Stemming — Удаление суффиксов из слов для приведения их к базовой форме. Например, «running» преобразуется в «run» [4].
- Lemmatization — Преобразование слов в их базовую форму с учетом контекста. Например, «better» преобразуется в «good» [4].

Данный принцип подходит только для статичных баз так как требует постоянной обработки текста.

3. Взвешивание термов.
- TF-IDF — статистическая мера, используемая для оценки важности слова в документе относительно коллекции документов [5].
- BM25 — Улучшенная модель взвешивания, которая учитывает длину документа и насыщение частоты терма [5].

Данные алгоритмы хорошо подходят для ранжирования текстовых документов.

4. Распределенные системы и параллелизм.
- Хранение и обработка индексов на нескольких серверах для повышения масштабируемости и отказоустойчивости (например, использование Apache Hadoop или Elasticsearch) [6].
- Использование многопоточности и параллельных вычислений для ускорения процесса индексации и поиска [6].

Данные методы касаются только физической реализации работы алгоритма и используются в основном на высоконагруженных распределенных базах данных.

5. Обработка естественного языка
- Named Entity Recognition (NER): Выделение именованных сущностей (имена, даты, места и т. д.) для улучшения поиска по контексту [7].
- Семантический анализ: Использование семантических моделей, таких как Word2Vec, GloVe или BERT, для учета смысловых связей между словами [7].

Данные алгоритмы делают поиск более схожим с человеческим, и являются наиболее популярными в поисковых машинах.

Обработка естественного языка считается наиболее продвинутой технологией улучшения инвертированного индекса за последние годы, что привлекает множество исследований на данную тематику. Например, возможность использования терм-документной матрицы на основе функции DGH для оценки патентов [8]. Или использование структуры множественных индексных векторов, где Word2Vec используется для генерации нескольких типов индексов под каждую выделенную

семантическую группу [9]. Но, несмотря на результаты этих исследований, вектор всё больше склоняется к использованию искусственного интеллекта в поисковых машинах [10].

Результаты

Исходя из вышеперечисленного можно сделать вывод об актуальности адаптирования алгоритма инвертированного индекса под совместную работу с нейронными языковыми моделями.

Примером такой интеграций может служить построение инвертированного индекса на основе эмбеддингов, где эмбеддинг — это вектор значений каждое из которых обозначает связь с семантическим термом, как это представлено на рисунке 1.

Индексы строятся по критерию связи между векторами. Таким образом, поиск может выдать документы, очень связанные по тематике, но не имеющие одинаковых слов. Связь тематик можно настраивать с помощью подстройки модели или использовать не все значения векторов для определения связи между словами.

Заключение

В данной статье была проанализирована тенденция изменений способов использования алгоритма инвертированного индекса, показана актуальность использования алгоритма совместно с нейросетевыми языковыми моделями, представлен пример построения инвертированного индекса по эмбеддингам сети.

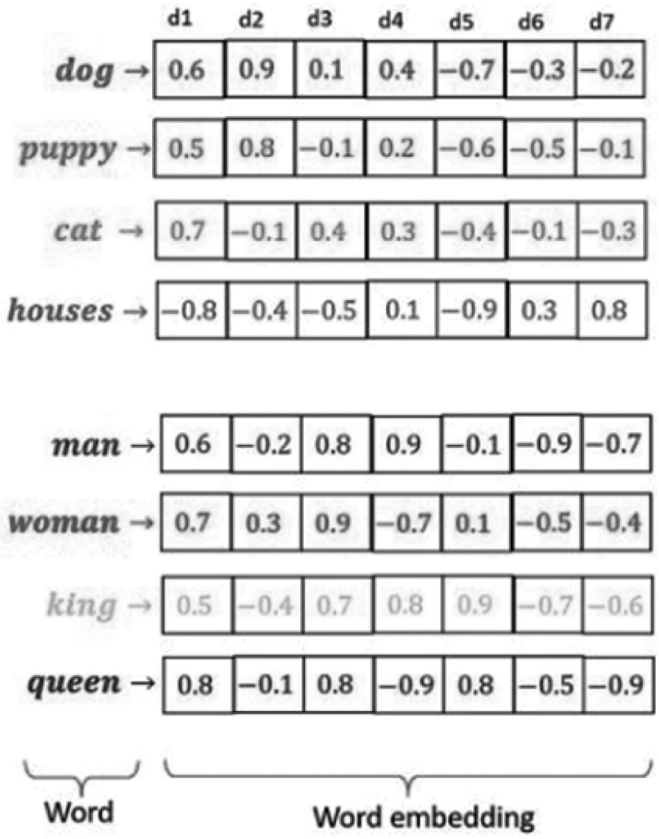

Рис. 1. Векторные представления эмбеддингов

ЛИТЕРАТУРА

1. Salton G. Automatic information organization and retrieval. — 1968.
2. Андрианов И.А., Корякин Д.С. РАЗРАБОТКА НОВЫХ ВИДОВ ИНДЕКСОВ ДЛЯ СУБД POSTGRESQL С ИСПОЛЬЗОВАНИЕМ ПРОГРАММНОГО ИНТЕРФЕЙСА СЕРВЕРА // СОВРЕМЕННЫЕ ИННОВАЦИИ В НАУКЕ И ТЕХНИКЕ. — 2020. — С. 20–24.
3. Pibiri G.E., Venturini R. Techniques for inverted index compression //ACM Computing Surveys (CSUR). — 2020. — Т. 53. — №. 6. — С. 1–36.
4. Desai D. et al. A comparative study of information retrieval models for short document summaries //Computer Networks and Inventive Communication Technologies: Proceedings of Fourth ICCNCT 2021. — Springer Singapore, 2022. — С. 547–562.
5. Marwah D., Beel J. Term-recency for TF-IDF, BM25 and USE term weighting //Proceedings of the 8th International Workshop on Mining Scientific Publications. — 2020. — С. 36–41.
6. Закиров М.З. Онтология больших данных информационных систем для оценки научного уровня развития регионов Российской Федерации //Искусственные общества. — 2020. — Т. 15. — №. 3. — С. 6–6.
7. Невзорова О.А., Хакимуллин Р.Р., Идрисов И.И. Цифровая научная платформа «Агрегатор неструктурированных геолого-промысловых данных»: архитектура и базовые модели извлечения данных //Георесурсы. — 2024. — Т. 25. — №. 4. — С. 149–162.
8. Коробкин Д.М. Метод формирования критериальных оценок морфологических признаков технических систем //Моделирование, оптимизация и информационные технологии. — 2020. — Т. 8. — №. 4. — С. 23–24.
9. Chen Y. et al. OneSparse: A Unified System for Multi-index Vector Search //Companion Proceedings of the ACM on Web Conference 2024. — 2024. — С. 393–402.
10. Барщевский Е.Г. Использование искусственного интеллекта //Восточно-Европейский научный журнал. — 2023. — №. 3–2 (88). — С. 56–58.

Журнал «Современная наука: актуальные проблемы теории и практики»